Vektortér alapú szemantikai szóhasonlósági vizsgálatok

Tóth Ágoston

Debreceni Egyetem, Angol-Amerikai Intézet
4010 Debrecen, Pf. 73.
toth.agoston@arts.unideb.hu

Kivonat: A bemutatott kísérletben kiválasztott szavakat a környezetükben előforduló szavak gyakorisági adataiból képzett vektorokkal reprezentáljuk, a vektorok összehasonlításával pedig a szavak szemantikai hasonlóságára következtetünk. A kísérleti rendszer egy feleletválasztásos feladatot old meg, melyben 30 célszó mindegyikéhez automatikusan kiválasztjuk a hozzá leghasonlóbb szót. A vizsgálandó szavak listáján 15 szemantikailag motivált párt találunk, köztük ellentéteket, szinonimákat és alá-/fölérendelt szavakat; kimenetként mindegyik szó párját vártuk visszakapni. A helyes választ a rendszernek mind a 30 szóhoz összesen 100 potenciális jelölt közül kellett kiválasztania. A pontosság maximális értéke (20 millió szavas korpusz feldolgozása után) 79% volt. A vektorokat a Magyar Webkorpuszból vett, annotációt nem tartalmazó szövegek segítségével állítottam elő, a vektorok kiszámítását és összehasonlítását saját fejlesztésű programmal végeztem.

1 Bevezetés

A szavak előfordulási gyakoriságára vonatkozó megfigyelések az ember és a gép által is könnyen gyűjthető adatok, melyek önmagukban is megalapozzák szemantikai jellegű feladatok megoldását.

Az első vektortér alapú szemantikai eredmények az információkeresés területén születtek (l. pl. [7]). Egy dokumentum a benne előforduló szavak gyakorisági adataival jellemezhető, melyekből (dokumentumokra jellemző) vektorokat hozunk létre. Ezáltal egyrészt a dokumentumok egymással összehasonlíthatók, másrészt az információkereséshez használt aktuális keresőkifejezésből ugyanilyen módszerrel létrehozott szógyakorisági vektorral a már meglévő vektorokat összehasonlítva a releváns dokumentumok megtalálhatók.

Szintén konstruálható olyan rendszer, amely nem dokumentumok, hanem szavak hasonlóságának mérését teszi lehetővé (l. pl. [3] és [6]). Ebben az esetben kiválasztott célszavakat olyan vektorokkal reprezentálunk, amelyek a környezetükben előforduló szavak gyakoriságát tükrözik. Az így kapott környezetvektorok összehasonlításával (pl. távolságuk meghatározásával) mérjük a szavak hasonlóságát, amelyet – a disztribúciós hipotézis [8] értelmében – szemantikai megfigyelésnek tekintünk. A vektorokat a környezetszavak által meghatározott sokdimenziós térben egyszerűen összehasonlíthatjuk úgy, hogy az origóból a vektorok által kijelölt pontok távolságát mérjük, vagy a

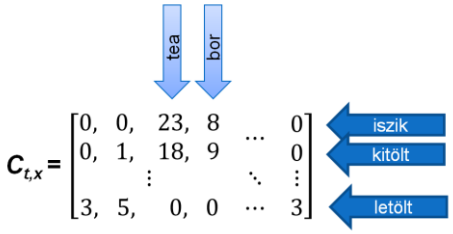
vektorok hajlásszögét állapítjuk meg. Az eljárás a szójelentés egy speciális közelítésének egyfajta geometriai modellezését jelenti.

Munkám egy olyan kísérletet mutat be, melyhez saját JAVA-alkalmazást fejlesztettem, mely nagyméretű korpuszokból automatikusan felépít előre meghatározott dimenzióval rendelkező vektortereket, és létrehozza a kijelölt szavakat jellemző vektorokat, amelyeket végül össze is hasonlít egy feladat megoldása során.

A cikk felépítése a következő: először bemutatom a szóhasonlósági kísérletemben használt rendszer felépítését a betanítás során használt paraméterek megadásával, majd leírom a kísérletben végrehajtott szemantikai feladatot, és értékelem a rendszer teljesítményét.

2 A kísérleti rendszer felépítése

Első lépésként egy mátrixot hozunk létre, melynek soraiban egy-egy *célszó* ábrázolását állítjuk elő (ezek megfelelnek a bevezetőben említett környezetvektoroknak), az oszlopok pedig egy-egy *környezetszónak* a célszavak környezetében megfigyelt előfordulási gyakoriságát reprezentálják. A mátrix egy eleme azt mutatja meg, hogy az adott célszó környezetében a feldolgozott korpuszban összesen hányszor fordul elő az adott pozícióhoz tartozó környezetszó.



1. ábra. Szó/környezet mátrix (t=*target*, x=*context*).

A mátrix sorait egy-egy környezetvektorként értelmezzük, amely a célszó és a környezetszavak közötti szintagmatikus kapcsolatokat ábrázolja. Például az 1-4 mondatok feldolgozása során az *ittam* szó környezetvektorában – egy 1+1 szavas szimmetrikus mozgó ablakot használva a környezet megfigyelésére – növelni fogjuk a következő szavaknak megfelelő vektorelemek értékét: *szóval, kávét, nem, teát* és *a*. Nagyobb, 2+2 szavas ablak esetén az *ittam* szó környezetvektorát befolyásolni fogják a *borból* és a *sörömet* szavak is. A vektorelem értéke arányos a célszó és az adott vektorelemnek megfelelő környezetszó közös előfordulásainak számával.

1. Szóval *ittam* kávét.
2. Nem *ittam* teát.
3. *Ittam* a borból.
4. *Ittam* a sörömet.

A környezetvektorok ábrázolásához szükséges vektorterek általában nagyon sok dimenzióval rendelkeznek, hiszen alapesetben minden, a jellemzett szavak környezetében előforduló környezetszó növeli a vektortér dimenzióját, amit utólag csökkenthetünk kezelhető méretűre. Jelen kísérletsorban elkerülöm a dimenzióredukciót azzal, hogy kizárólag a leggyakoribb (8-14 ezer) szót veszem figyelembe az ábrázolandó célszavak környezetében, a vektorok összehasonlítását pedig olyan egyszerű eszközökkel végzem, ami ilyen dimenziószám esetén is jól használható és gyors.

A célszavakat jellemző környezetvektorokat nem „nyers" formában (frekvenciaadatokkal) használtam fel, hanem belőlük a cél- és környezetszavakra pozitív pontszerű kölcsönös információt (pPMI) számoltam [2], ezzel mérve a két szó együttes előfordulásának valószínűségét azok *külön* történő előfordulásához képest.

Végül a pPMI értékeket tartalmazó vektorok összehasonlítása során a hajlásszögükből (α) számolt $\cos \alpha$ értékkel kaptam meg a célszavak hasonlóságát (vö. [5]). Előfeltevésünk szerint ez szemantikailag interpretálható mérték.

A hasonló kísérletek egyik fontos és általában hosszas munkával kikísérletezhető momentuma a lehetséges paraméterek megfelelő beállítása. Számos ilyen paraméter létezik a fent leírt, kifejezetten a rendszer felépítésére vonatkozó paramétereken kívül is. Ebben a kísérletben annotáció nélküli korpuszt használtam, tokenizálás és lemmatizáció nélkül, stopszavak használatát mellőzve (tehát a funkciószavakra vonatkozó gyakorisági adatok is megjelentek a környezetvektorokban, ami a pPMI vektorok és a hajlásszög alapú összehasonlítás miatt elvileg nyereséges döntés). A vektorok előállítása során a mozgóablak mérete 1+1 szó volt (bal és jobb oldalon 1-1 közvetlen szomszéd). Elsődleges célom a paraméterek beállítása során az angol nyelvre vonatkozó szakirodalmi adatok alkalmazhatóságának (elsősorban [1] alapján) kipróbálása volt a magyar nyelv feldolgozásában. Ebben a konkrét kísérletben a magyar és az angol nyelv közötti különbségek (gondolva itt elsősorban a nagyon különböző alaktani alrendszerekre) nem jelentettek problémát; ezzel együtt, bizonyos paraméterek eltérő beállításának a vizsgálata (pl. lemmatizáció használata) a későbbiekben szükséges lehet.

3 A szemantikai feladat, a rendszer pontossága

A vektortér alapú szemantikai rendszer tesztelésének módszertana egy további fontos kérdés, amire a nemzetközi szakirodalomban legalább 4 különböző eljárást találunk [1]:

- „TOEFL-teszt": feleletválasztós teszt, melyben néhány alternatíva közül kell automatikusan kiválasztani a megadott szóhoz jelentésben legközelebb állót;
- távolság összehasonlítása: ez is egy feleletválasztásos feladat, melyben adott célszavakhoz automatikusan kiválasztjuk a hozzá legközelebb álló szót; a választási lehetőségek tartalmaznak véletlenszerűen kijelölt szavakat a célszavak közül, valamint a vizsgált célszó egy előre kijelölt és célszavak közé felvett szemantikai párját (pl. szinonimáját, ellentétét, stb.), amit helyes kimenetként várunk;

- szemantikai osztályozás (előre kijelölt kategóriákba, pl. gyümölcsök, fegyverek, stb.);
- szófaji és mondattani klaszterezés.

Az itt bemutatott kísérleti rendszerben megoldandó feladatként egy távolság-összehasonlítási vizsgálatot választottam, amihez 30 célszót használtam. Ezek 15 szemantikailag motivált párt alkottak: voltak köztük szinonimák (pl. *egész–teljes*, *fut–rohan*, *néz–figyel*), ellentétek (*fekete–fehér*, *régi–új*, *ki–be*) és hiponimák/hiperonimák (alá-/fölérendelt szavak, avagy specifikusabb/általánosabb szavak, pl. *alma–gyümölcs*, *labdarúgás–sport*, *szekrény–bútor, kutya–állat*), egyforma számban. A figyelt szavak ilyen megadása azt biztosította, hogy mindegyik szóhoz volt egy „legközelebbi szó”, amely a rendszer által visszaadandó elvárt kimenet volt. A szavak kiválasztásánál a szófaji változatosságról gondoskodtam.

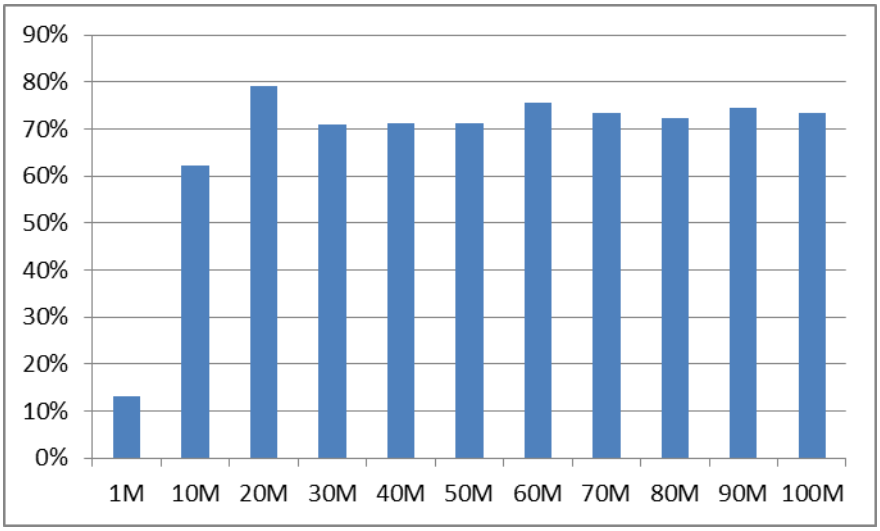
1. táblázat: A kísérlethez kiválasztott szavak.

Célszó	***Várt kimenet***
fekete	*fehér*
fehér	*fekete*
régi	*új*
új	*régi*
fent	*lent*
lent	*fent*
ki	*be*
be	*ki*
jó	*rossz*
rossz	*jó*
legmagasabb	*legnagyobb*
legnagyobb	*legmagasabb*
egész	*teljes*
teljes	*egész*
tép	*szakít*
szakít	*tép*
néz	*figyel*
figyel	*néz*
fut	*rohan*
rohan	*fut*
alma	*gyümölcs*
gyümölcs	*alma*
szekrény	*bútor*
bútor	*szekrény*
kutya	*állat*
állat	*kutya*
labdarúgás	*sport*
sport	*labdarúgás*
dollár	*deviza*
deviza	*dollár*

A helyes kimenetet a rendszernek mind a 30 szóhoz összesen *100 potenciális jelölt közül kellett kiválasztania*: a 100 alternatíva tartalmazta az eleve vizsgált 30 szót, valamint 70 olyan szót, amit a Magyar Webkorpusz [4] első 1000 leggyakoribb szavából választott a program véletlenszerűen. (Ilyen módon előfordulhat, hogy az opciók közé bekerül egy vagy több olyan szó, amely egy célszóhoz szemantikailag kapcsolódik. Ezt kizárni nem tudtam, de lent megadom a rendszer pontosságát arra az esetre is, amikor a 70 véletlenszerűen kiválasztott szó nem szerepelt a választható alternatívák között.) A véletlen elem miatt a futtatást többször megismételtem, és az eredményeket átlagoltam. A környezetvektorok kiszámítására a Magyar Webkorpuszból vett 100 millió szavas (annotáció nélküli) részkorpuszt használtam.

A random baseline pontosság 1% volt. A fedést ebben a tesztelési módszertanban 100%-on tartjuk: a feleletválasztás kikényszerített jellegű.

A pontosság 1 millió szó feldolgozása után átlagosan 13% volt, de ekkor még volt olyan célszó a 30 közül, ami a rendszer által figyelt környezetszavak (a Webkorpusz 14000 leggyakoribb szava) mellett még egyáltalán nem fordult elő a korpuszban. 10 millió szó után a pontosság 62%, 20 millió szónál 79% volt (baseline: 1%); ezután már nem javult a pontosság, egészen 100 millió szóig vizsgálva. A feldolgozott szavak száma a 2. ábrán látható módon befolyásolta a pontosságot.



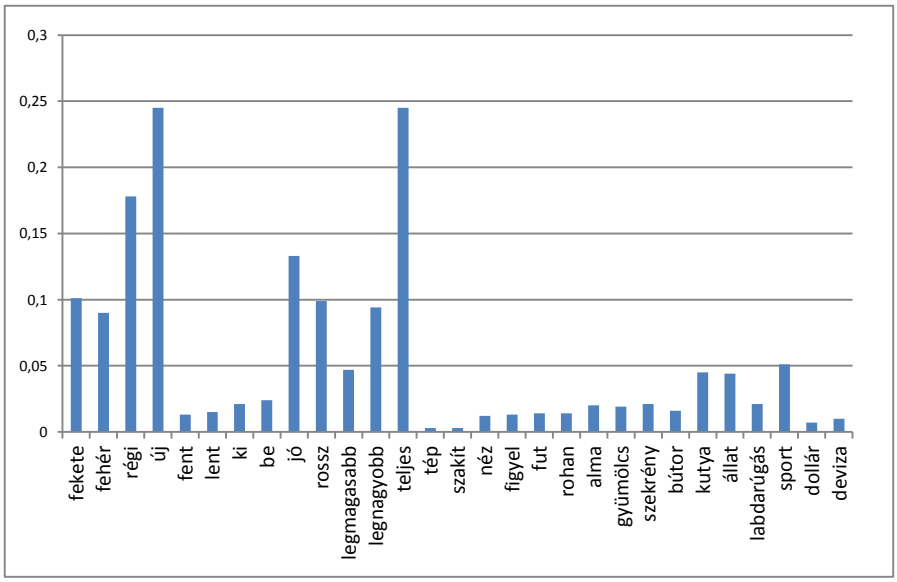
2. ábra. Pontosság változása a korpuszméret (millió szó) függvényében.

A választási lehetőségeknek a redukálása az eredeti 30 szóra javulást hozott (87% maximális pontosság 50 és 60 millió szavas korpuszméreteknél, 3%-os random baseline mellett). A választási lehetőségek 100-ról 250-re növelése a pontosságot csak enyhén, 77%-ra csökkentette (random baseline=0,4%).

A szakirodalomban elterjedt az, hogy a választási lehetőségek számát kifejezetten alacsony szinten tartják, így pl. 10 szóból választva (10% baseline mellett) elérhető 90% feletti pontosság is.

A kísérletbe bevont környezetszavak számát 8 és 14 ezer között vizsgáltam. Ennek a paraméternek a növelése marginális, de mérhető változást okozott (a környezetszavak számának emelése a pontosságot növelte, az elért növekedés néhány százalékos volt).

A számszerűsíthető eredmények mellett érdekes volt azon esetek vizsgálata, amikor egy adott szóhoz nem az elvárt kimenetet, hanem egy másik szót találtunk leghasonlóbbnak. A megfigyelt esetek egy része szemantikailag is értelmezhető volt. Ilyen például a *kutya→ember* és *állat→ember* asszociációk (*kutya↔ állat* helyett) abban az esetben, amikor a véletlenszerűen kiválasztott opciók között az *ember* szó is megjelent. Szintén a véletlen elemnek köszönhető probléma volt, amikor a *legmagasabb* szó párjának keresése közben a lehetséges válaszok közé bekerülő *magas* szó elnyomta az előre kijelölt párt (*legnagyobb*), ami tulajdonképpen nem is hiba, azonban az itt alkalmazott kiértékelési módszertanban a pontosság csökkenéséhez vezet. Természetesen arra is volt példa, hogy az algoritmus által visszaadott asszociáció szemantikailag motiválatlannak tűnő zaj volt, pl. *egész→új* megfeleltetés a kimenetként remélt *egész→teljes* helyett. A 3. ábra ezt az esetet mutatja be, szemléltetve az *egész* szó hasonlóságát az első 30 célszóhoz (a véletlenszerűen választott 70 szó hasonlósági értékeit itt helyhiány miatt nem ábrázoltam).



3. ábra. Célszavak hasonlósága az *egész* szóhoz (1=maximális hasonlóság).

Szemantikailag nem értelmezhető zaj esetén általánosnak volt mondható a 3. ábrán látható jelenség: a várt kimenetnek (ebben az esetben: a *teljes* szónak) és a zajnak (itt: *új*) a célszótól (*egész*) való távolsága nagyon hasonló volt, a harmadik, negyedik stb. helyezett szó jócskán lemaradva követte őket.

Általános tendenciaként megfigyelhető volt, hogy az alá-/fölérendelt szavaknál volt a legnagyobb a pontosság, ettől elmaradt az ellentétek és a szinonimák kezelése.

Munkám távlati célja a vektortér alapú számítógépes nyelvészeti megközelítés szisztematikus szemantikai vizsgálata, hiszen – miközben alkalmazásokban már megjelentek ezek az eszközök, és a velük kapcsolatos tapasztalatok egyre gyűlnek –, lexikai szemantikai szempontból az ilyen eljárásokat nem értékelték még mélyrehatóan. A számítógépes eszköz kifejlesztése és kipróbálása az itt bemutatott módon az ehhez szükséges első lépés volt.

Köszönetnyilvánítás

A cikk elkészítését részben az OTKA K 72983 számú kutatási projekt, részben pedig a TÁMOP 4.2.1./B-09/1/KONV-2010-0007 számú projekt támogatta. A TÁMOP projekt az Európai Unió támogatásával, az Európai Szociális Alap társfinanszírozásával valósult meg.

Bibliográfia

1. Bullinaria, J.A., Levy, J.P.: Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior Research Methods, Vol. 39 (2007) 510–526
2. Church, K.W., Hanks, P.: Word association norms, mutual information, and lexicography. Computational Linguistics, Vol. 16 (1990) 22–29
3. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society for Information Science (JASIS), Vol. 41 (1990) 391–407
4. Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V.: Creating open language resources for Hungarian. In: Proceedings of the 4th international conference on Language Resources and Evaluation (LREC2004) (2004)
5. Landauer, T.K., Dumais, S.T.: A solution to Plato's problem: The Latent Semantic Analysis theory of acquisition, induction and representation of knowledge. Psychological Review, Vol. 104 (1997) 211–240
6. Lund, K., Burgess, C.: Producing high-dimensional semantic spaces from lexical cooccurrence. Behavior Research Methods, Instruments & Computers, Vol. 28 (1996) 203–208
7. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Communications of the ACM, Vol. 18, No. 11 (1975) 613–620
8. Turney, P.D., Pantel, P.: From frequency to meaning: Vector space models of semantics. Journal of Artificial Intelligence Research (JAIR), Vol. 37 (2010) 141–188